CHAPTER FIVE: EXAMPLES

Sometimes, providing an example is more worthwhile than a thousand words of advice. Because we believe that is the case with evaluation, we have provided a number of examples (both good and bad) of different evaluation efforts. These examples build on information from the previous chapters. The specific examples we present are fictitious, but are based on studies that could and have been done.

Example 1: Evaluation of an Inservice Program for Elementary Science Teachers

This example illustrates the:

- **Use of formative evaluation for measuring implementation**
- **Importance of involving a skilled evaluator early in the project**
- **Use of both qualitative and quantitative approaches**
- **Use of multi-informants and data collection techniques**
- **Ability to adapt a design based on new information.**

NSF funded the Center for Professional Enhancement for a 3 year program of in-service education for elementary science teachers. The purpose of the project was to introduce the teachers to some of the new approaches to elementary science instruction and to assist them in applying these techniques in their classrooms. Teachers were selected for participation based on the nomination of a supervisor (usually a principal), their written essay, and a commitment both to attend regular and summer sessions and try out the new approaches in their classrooms.

The program consisted of 1 year of training with follow-up workshops after the first year. A mixture of teaching strategies was used. These included: lecture sessions, hands-on experiences in using techniques identified as being promising, visits to classrooms taught by model teachers, and peer coaching.

Before the project was even funded, the Principal investigator hired an evaluator to participate in the program. The evaluator was one who had considerable experience in studies of elementary science teacher training and was aware of the new directions in which elementary science is proceeding. The Principal investigator had considered hiring an elementary science teacher who had been surplused to fill this role, but decided against it because of the belief that a more skilled evaluator would benefit the project more. The evaluator and the Principal investigator discussed the goals of the project, the plan for meeting these goals, and the questions that needed to be explored. Considerable time was devoted to clarifying the kinds of information needed to make sure the training was functioning as intended. The Principal investigator was very interested in studying program implementation as early as possible to make sure everything was

"on track." The evaluator also talked to the trainers to understand their information needs and understand more fully the kinds of data that would help them do their jobs as well as possible. Both the Principal investigator and the trainers expressed a strong interest in knowing the extent to which what was learned was, in fact, being transferred to the classroom.

The evaluator began observing the training sessions on an intermittent basis almost as soon as they started. Although no formal observation system was used, she wrote a brief narrative summary after each observation session detailing the focus of the session, the strategies discussed, and the involvement/engagement level of the participants. After 6 months, she administered a questionnaire to the participants which addressed a wide range of issues, including the adequacy of the training and the extent to which it was used and useful in their own classrooms. She also interviewed the trainers to get their reactions to the project and to hear any new concerns that may have arisen. She had planned to interview other teachers who worked with the participants to assess their awareness of the new teaching techniques (it was hoped that the training would have a "spill-over" effect on others in the school), but this idea was abandoned as being premature after a preliminary review of the teacher questionnaires.

The findings proved to be very useful and the Principal investigator was pleased with the feedback from this early investment. The observational summaries indicated that the training sessions themselves were highly effective. Even during the lecture sessions, participants were engaged and very attentive. The quality of interaction and discussion during the hands-on sessions was very good, with the participants frequently going beyond the demonstration tasks and inventing their own alternatives.

By and large, the interviews with the trainers complemented the observational data. Trainers were pleased at how smoothly their classes were going and at the high level of engagement shown by the participants. They did, however, have some problems with the model teachers and coordinating demonstrations by them with material covered in the other training sessions. Because these teachers were teaching in regular school situations, the needs of the project and the regular classroom too frequently came into conflict.

The data from the participant questionnaire were less positive. While participants had high praise for the training they were receiving, they were somewhat less enthusiastic about the project overall. While they had initially been very pleased with the opportunity to participate, they were finding that the time they spent away from the classroom was interfering with their jobs as teachers. Further, they were unable to apply what they learned to their own classes because they lacked the supplies and materials necessary. The support that had been provided for the traditional science teaching in which they previously had engaged did not meet the needs of the new lessons to which they were being exposed.

Based on these findings, several changes were made in the project. First, video-taped demonstrations were substituted for visits to model teachers during the regular school year. Arrangements were made for demonstrations of model teaching during the summer at a nearby laboratory school that had a summer session. Selection criteria for teachers were also changed so as to require some in-kind support from their institutions. The Principal or someone in the central office had to agree to provide the supplies and materials needed for instructing in the new ways, if such materials were not already available. The Principal investigator also reallocated some of the project funds to provide more materials for the teachers. A number of lessons were designated as ones in which all needed materials would be provided to the participants for use in their classes. Time was also set aside during the training sessions to discuss attempts to apply the new strategies to the classroom. Both successful and unsuccessful applications were considered.

The second year evaluation showed that these changes were having a positive effect. While the second year participants still had a certain level of frustration at being out of their classrooms, their ability to bring back and try out new strategies reduced this frustration greatly. The sharing sessions at which application attempts were discussed became favorites of both the participants and the trainers. The former gained important insights into ways of transferring their skills; the latter gleaned many tips to pass on to the next year's participants.

Example 2: Evaluation of an Integrated Learning System for the Teaching of Mathematics

NSF funded Jones University, along with the Smith School District to conduct a study of the efficacy of the

Boston Integrated Learning System (ILS). The Smith School District, once considered a very fine school system, had over the last several decades, fallen on hard times. Budget cuts, population shifts, and a national economy which has been sliding, combined to give Smith new challenges that it had never before faced. The results were discouraging. Not only were test scores on the decline, but absenteeism was high, and even when in school, the students frequently missed class or were disruptive.

The Boston ILS was selected for implementation in this district because of both its ability to tailor instruction to individual needs and its motivational characteristics. It was also hoped that with an ILS like Boston in place, teachers would be able to spend more one-on-one time with each student, without reducing the quality of instruction received by the group.

This example illustrates the:

- **Problems that can arise when the potential needs of critical stakeholders are not considered**
- **Limitations of relying on a single measure of program impact**
- **Need to provide for formative progress evaluation**
- **Misinterpretations that result from failure to appropriately disaggregate results.**

The Boston ILS provides the hardware and software needed to assist students in elementary mathematics. The ILS combines teaching modules, testing modules, and a reporting component intended to provide an individualized learning experience. It has high quality graphics and an audio component.

Seven schools were selected for the project. The schools were among the most needy in the district, defined in terms of student test scores and free lunch counts. All students in these schools participated.

The goal of the project, as written in the proposal to NSF, was to increase test scores in mathematics. The key indicator of performance was defined as scores on the Schmata Test of Basic Skills, a norm referenced test given every other year to the students.

The study was initiated in the fall of 1989. Because the evaluation design seemed to the Principal Investigator to be very straightforward—pre-post test performance on the Schmata test, no allowance was made for an evaluator at project onset. Rather, the Principal Investigator intended to rely on the normal test reporting of the school district as the means of acquiring evaluative data. He also included some funds for reanalysis of these data by his colleagues at Jones University. The NSF Program Officer queried the Principal Investigator about the scope of the evaluation, raising the question of whether or not it would meet all stakeholders' needs. She asked, specifically, whether or not all relevant parties at the school district had been consulted before

the proposal was submitted. The Principal Investigator said that he had talked to the Director of Curriculum and they were in agreement with regard to the evaluation. He said, however, that he would revisit the questions with a broader group of policymakers at the school level and amend his proposal, if necessary. In the rush of next steps, this consultation fell by the wayside.

Ten months after the project was initiated, the school district faced another budget crisis. The Board asked for evidence that the project was successful, threatening to cut back the in-kind funds that had been allocated for teacher training and planning time to the project schools. The Principal Investigator was able to give his impressions of how the program was working and several teachers also offered their support. In addition, however, the Board received a number of phone calls from other teachers saying that Boston placed too much of a burden on them and the benefit to students was negligible.

Fearing loss of support from the school system, the Principal Investigator called upon some of his colleagues for advice on what to do. Fortunately, the University had on staff some strong educational evaluators. After review of the project and the data available, they came up with an evaluation scheme. Recognizing the fact that the Schmata test results would not be available for more than a year, they gathered some interim measures of program impact. First, they looked at test performance on the assessment modules provided by Boston. Analysis of these data showed that, overall, students were making steady progress and seemed to be retaining the skills learned as measured by the retention tests. (While there was no way to compare these students' performance with that of students given traditional instruction, the analysis at least provided some promise of success.) Second, they looked at data in other areas to see whether an impact could be posited. The areas they selected were attendance and referrals for behavioral incidents. These data showed that, overall, attendance had increased and behavioral incidents had decreased compared to the same time in previous years. Further, when the data for some individual students were examined, these same students showed increased attendance and fewer referrals for disturbance than they had in the past. Taken together, these findings provided weight to the claims of the Principal Investigator and the project continued to receive support from the school district.

Six months later, the evaluators returned to these data and did some additional analyses, disaggregating the results by grade, gender, race, and English-language proficiency. The evaluators found the picture of success which emerged from the overall data did not hold true for each subgroup of students participating. Specifically, they found that both the progress and behavioral data showed striking differences between English and limited-English speaking students, with the limited English speaking students failing to do as well. Despite the fact that the latter were receiving the district's special language supports and were assigned to the regular classroom for their instruction, these students were clearly failing to profit from Boston ILS. Unfortunately, this information was not obtained until after the students had spent a full school year in the project andhad started a second year of participation.

Example 3: The Evaluation Of A Special Program For Gifted Minority Students

Project REACH FOR THE STARS, a project aimed at identifying and supporting gifted minority students in math and science, was funded by NSF under the Comprehensive Regional Center for Minorities Program. This project aimed at identifying talented minority students at the end of 8th grade and encouraging their participation in mathematics and science courses for the duration of their high school career and beyond. It included a mentoring component, Saturday morning and summer enrichment sessions, and support groups.

Students were identified for participation based on test scores, grades, and motivation. Two subgroups were created. Subgroup I, the majority of students (75%), had the highest test scores and at least a "B "average in math and science. The second subgroup, subgroup II (25%), consisted of students who were highly motivated to participate, but had not shown strong performance in the past. Since there were more students who qualified for the program and were interested than those who could be accepted, a lottery system was used to select individuals for participation. Those who were not selected were placed into a comparison group against which to measure the progress of students in subgroup I. Unfortunately, there was not a sufficiently large number of students who fell into subgroup II to allow for a similar procedure.

The evaluation used a wide variety of measures which were entered into a data base student-by-student.

Included were grades both in the target courses, math and science, and in other major academic subjects; test scores from end-of-semester exams; standardized tests; and, as relevant, the SAT, ACT, and College Board results. At the end of the 12th grade, data on post secondary applications and acceptances, as well as scholarships and other honors were obtained and added.

In addition, focus groups were conducted each year with participants in order to get students' reactions to the program both from an academic and a social perspective. Surveys were administered to the parents and teachers at the end of the second and fourth years. Finally, a follow-up survey was sent to graduates (both the participants and the subgroup I comparisons students) one year after they had graduated in order to find out what they were doing, how well they were doing, and what they thought, in retrospect of the special program in which they had participated.

This example illustrates:

- **A Summative Evaluation built on progress data collected annually**
- **The use of both survey and case study methodology**
- **The use of multiple data sources**
- **The problems of interpreting findings without a comparison group**
- **The consideration of timeliness in the production of a report.**

A substudy, which was turned into a dissertation by one part-time researcher, looked closely at the experiences of five students differing in gender, race, and family structure. These case studies were used to provide a thick description of program experiences and student/ family, and staff reactions.

On an annual basis, the data were analyzed for the program participants overall—the comparison group students and for the participating students by subgroups. These annual analysis were used to monitor the progress of the students and to pinpoint individual student problems as they arose. The student focus groups also provided important input for modifications in the project, which fine-tuned the approach.

A final report built on the data collected annually providing an overall summary for the four years of project participation. The only new data added in the fourth year that was not collected previously was the information on honors, awards, and post-secondary acceptances. Because of the careful job that had been done documenting progress along the way, the production of a final report was greatly simplified and there was little protest from any of the participants about the requests for data and the "burden" that it caused.

The final report showed that in general the program was a success. Students made good grades in the

target courses, continued to do well in the other courses, and were accepted into strong post-secondary institutions. There was a statistically significant difference between the grades and test scores of the program participants and the comparison group students. The data collected from the students' parents showed they had high expectations for their children and were convinced by their proven success in the program that they could and should aim high in the future. However, the parents of the nonparticipating students were quite similar in their responses. Where students from subgroup I dropped out or failed for other reasons to succeed, there was usually some extenuating circumstance relating to family or friends. Although the number was too small to be statistically significant overall, there was a tendency for comparison students to drop out more frequently for reasons related to school problems.

Although not all of the subgroup II students were successful, nearly half of them were so. Unfortunately, the analyses did not uncover any particular predictors of who from that group might succeed or fail. Some teachers felt that despite these students failure to attain success in absolute terms, they still performed better than they would have without the program. However, the lack of a comparison group for these students made it impossible to test out this hypothesis. The evaluator felt that understanding of these students could be enhanced by some further interviews or by the data that would result from the follow-up questionnaire and hoped to delay reporting until these data could be analyzed and the picture made more complete. However, the principal investigator felt that the report could not be delayed further without putting in jeopardy any chance of continued funding.

Example 4: Evaluation of a Summer Camp For Female High School Students

Project CAMP CRUSADE FOR WOMEN IN SCIENCE, a five-year project begun in 1987, is aimed at women in high school grades 9–11 and seeks to promote interest and involvement in the study of science. The goals are science-oriented high school and post high school course and activity choices on the part of camp participants which will ultimately lead them to pursue careers in the sciences.

This project, funded by NSF under the Young Scholars Program, currently has an all-woman

staff of two secondary science teachers, two undergraduate students majoring in science, and one college professor.

The project is being carried out by Hill College, a small midwestern institution located in the Mountain school district,a large district with 5 high schools. The participating college professor is the Principal Investigator.

Eligible applicants include all female students in grades 9–11 in the Mountain school district. All applicants are asked to complete a questionnaire which seeks information about previous courses taken, and includes a series of questions measuring attitudes, satisfaction, motivation, educational goals, and career goals. The Principal Investigator considered using Grade Point Average and test scores as selection criteria, but rejected this approach because of questions she had regarding how well they would predict performance on activities at the camp, and because she wanted Camp Crusade to provide opportunities for all women interested in science regardless of their previous attainments.

The summer camp provides opportunities for up to 35 participants to engage in a variety of activities such as lectures, experiments, field studies, films, and study work groups. From among 120 applicants, the 35 participants were randomly selected with equal selections from each grade.

The proposal to NSF included an evaluation component with a modest budget. It called for a Formative Evaluation to be conducted every two years. The project evaluator was a part-time instructor in the Department of Education at the Hill College with prior experience in educational research, but no evaluation experience. The evaluator planned to use an experimental design based on comparisons between the treatment group (camp participants) and a control group which consisted of a random sample of 50 non-accepted applicants. The evaluator attempted to match participants and controls by grade level, but given the small applicant pool this was difficult, since the great majority of applicants were 10th graders.

The first Formative Evaluation of the program took place in 1987. However, the initial evaluation was limited to an Implementation Evaluation that determined that the project was being conducted as planned. No Progress Evaluation was done to determine

whether the participants were moving towards meeting the project's goals. Stakeholders questioned the absence of progress information and wondered whether it was an oversight or an unwillingness to look at the issue.

To meet these concerns, the Principal Investigator and the evaluator decided to modify the evaluation design. They decided that an Implementation Evaluation would be conducted every other year, and a Progress Evaluation would be conducted yearly. The revised evaluation design called for data to be collected both at the end of the summer and during the next school year. Both qualitative and quantitative approaches were to be used to capture a wide variety of information on how the summer experience was affecting the participants. Specifically the data collection would include:

- Questionnaires
- Personal interviews
- Observations
- School records.

During the last two days of the camp, the evaluator conducted personal interviews and observed the work groups; on the last day she distributed a self-administered questionnaire to the participants, The interviews focused primarily on the camp experience while the self-administered questionnaire was essentially a replication of the questionnaire which all applicants had completed. The evaluator had planned to administer this questionnaire also to the control group but could not reach most students in the control group because of the timing of this survey (late summer, before the start of school) and this idea was abandoned.

Analyses of the initial information gathered on the experimental and control students showed that the groups were very similar. There were no differences between the two groups with respect to courses selected or motivation. Overall, the data from the experimental group collected following the camp experience appeared promising. The post-test questionnaire indicated that the camp attendees were more motivated than they had been to pursue advanced and elective high school course work in the sciences, had increased positive attitudes towards science, and were more

likely to consider pursuing future academic studies in the sciences than before they attended the camp. The interviews with the experimental group supported the findings from the questionnaires. They allowed the evaluator to add more qualitative data to the study and enhance the quantitative results.

The observational data suggested that the work group sessions were well-liked by the girls. There was a high level of interaction among the participants, as well as between the participants and the faculty. There was an especially high volume of contact between the campers and the undergraduate counselors. The students were observed asking numerous questions of the counselors regarding their college studies and the nature and difficulty of their programs. There were similar levels of interaction with the high school teachers, but interactions with the professor were constrained. Further observations will be necessary to determine why these interactions were so low, focusing on program design, personalities, etc. This issue was highlighted for further attention in the next Implementation Evaluation.

At the end of the second semester of the school year following the 1988 camp, a review of the school records indicated that 65% of the girls who attended the camp had registered for honors or advanced placement science classes, while only 45% of the control group students had done so. Also, 25% of the experimental group, and only 10% of the control group, chose to take a science-related course as their elective.

A follow-up questionnaire was mailed to both the experimental and control groups six months after the camp. The response rate was low for both groups—50% for the experimental group and 30% for the control group. The Principal Investigator recognized that follow-up phone calls were needed to increase the response rate and insure reliability of findings, but budget constraints prohibited this approach and these data were used in the evaluation although the evaluator was careful to point out their weakness.

The findings indicated that although the measures of attitudes and motivation for the experimental group had decreased slightly since the immediate post-test questionnaire was administered, the measures were significantly higher than those of the control group. Responses to questions about the

This example illustrates:

- **Ability to adapt a design based on new demands**
- **Use of control group for measuring project effects**
- **Utilizing an appropriate mix of data collection techniques - both quantitative and qualitative**
- **Failure to adequately plan for follow-up data collection procedures and correction for low response rates**
- **Over generalizing from one particular sample of participants to a whole population**
- **Treating data from a Progress Evaluation as if they were part of a Summative Evaluation**

camp itself indicated that the participants really enjoyed it and were happy about the new relationships they had established with other students and with the staff. Approximately 25% of the responses indicated that the camp participants desired more environmental science activities. Therefore, the Principal Investigator decided to accommodate this desire by replacing one of the lab activities that received very critical comments with an environmental science activity.

The report on the Progress Evaluation was very positive. Its overall conclusion was that this project had a positive effect and motivated girls to become more involved in science and oriented toward academic studies in scientific fields. In fact, the report included a recommendation that more camps like this should be established throughout the school district and more girls should be encouraged to apply.

The stakeholders who initially called for the Progress Evaluation greeted the findings and recommendation with mixed reactions. Supporters of the project were very enthusiastic about the findings. Critics questioned the conclusions citing the very high initial motivational levels of the experimental and control students and the low response rates in the follow-up survey. They finally agreed that the data were encouraging but that final decisions regarding program expansion needed to await additional data.